SLAUGHTER AND MAY/

March 2024

DATA SCRAPING AND COMPLIANCE: NO 'CLEARVIEW' (YET)?

Companies often rely on data scraped from publicly available sources, but what are the legal bases?

A version of this briefing first appeared in the Privacy Laws & Business UK Report, Issue 132 (March 2024)

Data scraping became somewhat notorious following the Cambridge Analytica scandal in 2018. It has continued to be controversial since then, leading to data privacy authorities (DPAs) from around the world publishing a joint statement last year setting out their concerns. However, despite this and DPAs such as the ICO taking specific action (e.g. in relation to Clearview AI Inc (Clearview)), the lawfulness of data scraping has not yet been settled. Now, with the hype around generative AI models, which typically require large quantities of data for training purposes, the debate has yet again come to the fore.

What is data scraping and what are the key legal challenges?

Data scraping or "web scraping" has various definitions but essentially means the process of gathering, copying or extracting information (including text, images and videos and therefore often personal data) from the internet with a view to storing that information and using it (or selling it for use). It can have many beneficial use cases such as for fraud or background checks or for training AI models.

However, data scraping gives rise to a number of concerns given the large volumes of data collected that can be used in ways with very real and substantial effects on people's lives. In addition, the process is often largely invisible, with individuals being unaware that details about their lives are collected and used in this way. Unsurprisingly, this makes data privacy compliance quite challenging.

In the UK, data scraping has attracted the attention of the ICO, as evidenced by its enforcement against Clearview and its "Generative AI consultation: the lawful basis for web scraping to train GenAI models" published on 15 January 2024 (GenAI Consultation), both of which provide useful insights into the ICO's approach to data scraping. However, the position is not settled, with the Clearview case currently on appeal to the Upper Tribunal (albeit not for reasons relating to data scraping and data privacy compliance) and the consultation is still open for comment. In the EU, data scraping is also on the radar of DPAs, with a number of them (including in Greece, France, Austria and Italy) also having taken action against Clearview.

Scraping personal data (and its subsequent use) touches on most aspects of the (UK) GDPR. For example, the process is likely to be a high-risk activity that needs to comply with the privacy by design principles and is an example of "invisible processing" that the ICO says requires a data processing impact assessment (DPIA). Challenges around data minimisation and purpose limitation will also need to be addressed, but in this article we focus on what we see as the key issues of legal processing grounds, transparency, fairness and lawfulness.

Processing ground under Article 6

A key issue is whether there is an appropriate processing ground under the UK GDPR for scraping personal data. In the Clearview case the ICO stated that the onus is on the controller to demonstrate it can rely on one or more lawful bases under Article 6. This was a significant issue for Clearview, with the ICO concluding in its monetary penalty notice (MPN) that Clearview did not have a processing ground it could rely upon. In fact, Clearview did not attempt to argue at the First Tier Tribunal (FTT) hearing that any of the grounds in Article 6 were met, possibly because the Italian DPA had already concluded that Clearview could not rely on the legitimate interests processing ground. Whilst not mentioned in the Clearview MPN or the subsequent FTT decision, an organisation's failure to demonstrate that it has met one of the processing grounds when carrying out data scraping could be treated as a breach of the GDPR's accountability obligations in of itself.

In its GenAI Consultation, the ICO states that "based on current practices, five of the six lawful bases are unlikely

to be available for training generative AI on web-scraped data", with the legitimate interests ground being the only possibility. This view seems to be mirrored by other DPAs, with the European Data Protection Board (EDPB) reporting that the Austrian DPA found that "Clearview AI could only have been covered by the legitimate interests ground in Article 6(1)(f) GDPR". This does, however, ignore the position of public bodies who are not able to rely on the legitimate interests ground, and since the ICO has usually shied away from such categorical statements, we envisage that its position may soften in its final guidance.

When considering the legitimate interests processing ground, a key challenge is how the data scraper can pass the balancing test when it may have limited insight into, or control over, how any model that is trained on its data is used and to what purpose by the end-user. The Austrian DPA considered this to be an issue in its action against Clearview, with it concluding that, due to serious privacy intrusion, the interests of the complainant clearly outweighed the purely commercial interests of Clearview.

This balancing exercise is one of the areas on which the ICO is requesting feedback. In what appears to be a pragmatic and innovation-friendly approach, the ICO suggests ways in which Gen AI developers can pass the legitimate interests balancing exercise, including by putting in place technical measures such as output filters, organisational controls over specific deployments, contractual protections and audit requirements where models are made available to third parties. However, how practical some of these measures would be is unclear, particularly in relation to the requirement to audit third parties' use of the model.

Transparency and fairness

The ICO set out in the Clearview MPN that individuals were not made aware of the processing by way of a privacy notice. They would only therefore become aware of the processing if they came across Clearview's website or read about the processing in the media. A similar point was made by the ICO in the Experian enforcement notice. A further concern the ICO had in relation to Experian, was that, for those that did receive a privacy notice, the content and its format (being heavily layered) did not permit the reader to understand how their data was being used.

It can be a challenge to explain complex processing activities like data scraping in a manner that strikes the right balance between detail and understandability. The ICO decision in Experian was appealed, and the First Tier Tribunal agreed that Experian was required to provide notices to data subjects whose personal data it obtained from public sources, but it was more pragmatic when it came to the content and format of the notice, arguing that there is a "tension between providing large amounts of information on the one hand with the aim of improving transparency and accessibility of information and on the other the resultant information overload." That said, this is currently being reconsidered in the ICO's appeal of the Experian decision to the Upper Tribunal, with the ICO's position being that the focus should not only be on the consequences of processing, but also on whether individuals would find it surprising and the fact that their rights are ineffective if processing is invisible.

Transparency and fairness are often interlinked - in Clearview's case, the ICO decided that the data collection was unfair because individuals would not expect their public personal data to be collected and used for the purpose of facial recognition by potentially a wide range of end users. However, the FTT was not required to examine the substantive findings of non-compliance against Clearview in this regard given it decided that the GDPR did not apply to the processing.

The fairness and transparency issues are fundamental challenges in the context of data scraping because it will always be near-impossible for any data scraper or its customers to bridge that knowledge gap with individuals, ensuring that the processing does not amount to "invisible processing". So, subject to where the Upper Tribunal lands on this in the Clearview case (if indeed it ultimately needs to consider these aspects), the ICO may take the same approach in future and fine organisations for fairness and transparency failings in the context of data scraping.

Lawfulness principle

To be lawful, not only must the data scraping comply with the wider requirements of the GDPR, but developers of GenAI must also ensure that their data scraping is not in breach of wider laws, such as intellectual property law or contract law.

Copyright

Much of the online content that an organisation might wish to scrape will be subject to copyright so, to be lawful, would need to benefit from an exemption or benefit from a licence.

However, it is far from clear that any of the exceptions to UK copyright apply in respect of most data scraping for commercial purposes. The most relevant exception is that for text and data analysis, but this only applies where such analysis is "for the sole purpose of research for a non-commercial purpose". A few years ago, the UK government did recommend expanding this to non-commercial purposes to encourage AI development in the

UK, but it has since shelved its proposal following significant outcry from the creative sector.

The alternative is to obtain licences from relevant rights holders, and some media organisations have started to grant licences which permit content to be collected and used to train AI models. However, it is often unfeasible to seek a licence from all relevant rights holders given the volume of data being scraped.

This makes it very difficult to undertake broad data scraping in the UK for commercial purposes without infringing copyright. Given the GDPR lawfulness requirement, a breach of copyright also results in the data scraping breaching the GDPR. Whilst the ICO obviously cannot change the law, it would be helpful if it could expressly state that it would not take enforcement action purely on the basis of a breach of copyright. Longer term, we hope that the UK government will take action to reach a position which appropriately balances the rights of the content owners with that of innovation.

Website terms of use terms

Most websites will have specific terms and conditions that users, including a data scrapers, agree to comply with when accessing them. Often, these Terms of Use include a provision that limits the purposes for which access can be gained. If personal data is scraped in breach of this purpose restriction, then technically, the scraper will act in breach of contract. This could potentially lead to lawfulness issues for both the data scraper and its customer on the basis that any processing in breach of contract might result in non-compliance with the GDPR principle.

A solution would of course be to ensure no data is scraped from websites that do not permit it, but there is a question whether this kind of manual assessment would be feasible if the scraper relies on the data of a large number of websites.

AI considerations

Any entity that will be subject to the EU AI Act will have to consider the ways in which this legislation may limit data scraping. In particular, the leaked draft of the AI Act prohibits any untargeted scraping of facial images to create facial recognition databases.

Closing thoughts

Whilst the challenges of data scraping are clear from the above, the value of its use-cases should not be underestimated; when used correctly and proportionately it can have many real advantages that benefit society as a whole, from spotting fraud to training AI models which can improve everyday life.

The ICO is clearly grappling with the question of how data scraping can be undertaken in a compliant manner, evidenced by its appeal of the FTT's Clearview decision and its Gen AI Consultation. The topic is receiving similar attention in the EU and elsewhere in the world, such as the US where there has been a string of cases.

Given the potential benefits of data scraping, we hope that the governments and regulators in the UK, the EU and elsewhere in the world will take steps to resolve the challenges in a pragmatic way, with appropriate balancing of the risks and benefits, so as to provide the necessary legal clarity to enable innovation.

The ICO states in its Gen AI Consultation that it is "moving fast to address any risks and enable organisations and the public to reap the benefits of generative AI." In our view, the ICO's approach is a step in the right direction, although it does not have the power itself to resolve all the challenges discussed above, and fits with the "pro innovation" approach taken in the new Data Protection and Digital Information Bill. Let's hope the UK government and others follow suit.

CONTACT

REBECCA COUSIN
PARTNER
T: 020 7090 4738
E: Rebecca.Cousin@slaughterandmay.com

LUCIE VAN GILS
ASSOCIATE
T: 020 7090 3560
E: Lucie.vanGils@Slaughterandmay.com

IAN RANSON
ASSOCIATE
T: 020 7090 3932
E: Ian.Ranson@slaughterandmay.com

London
T +44 (0)20 7600 1200
F +44 (0)20 7090 5000

Brussels
T +32 (0)2 737 94 00
F +32 (0)2 737 94 01

Hong Kong
T +852 2521 0551
F +852 2845 2125

Beijing
T +86 10 5965 0600
F +86 10 5965 0650

www.slaughterandmay.com

585024086